Interpretable brain age prediction using linear latent variable models of functional connectivity

Ricardo Pio Monti[1,7], Alex Gibberd[2], Sandipan Roy[3], Matthew Nunes[3], Romy Lorenz[4,5], Robert Leech[6], Takeshi Ogawa[8], Motoaki Kawanabe[7,8], Aapo Hyvärinen[9,10]

1 Gatsby Computational Neuroscience Unit, University College London, London, UK
2 Department of Mathematics & Statistics, Lancaster University, Bailrigg, UK
3 Department of Mathematical Sciences, University of Bath, Bath, UK
4 MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge, UK
5 Department of Psychology, Stanford University, Stanford, CA, USA
6 Centre for Neuroimaging Science, Kings College London, London, UK
7 RIKEN Center for Advanced Intelligence Project (AIP), Kyoto, Japan
8 Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International (ATR), Kyoto, Japan
9 Université Paris-Saclay, Inria, Palaiseau, France.
10 Department of Computer Science and HIIT, University of Helsinki, Helsinki, Finland

Abstract

Neuroimaging-driven prediction of brain age, defined as the predicted biological age of a subject using only brain imaging data, is an exciting avenue of research. In this work we seek to build models of brain age based on functional connectivity while prioritizing model interpretability and understanding. This way, the models serve to both provide accurate estimates of brain age as well as allow us to investigate changes in functional connectivity which occur during the ageing process. The methods proposed in this work consist of a two-step procedure: first, linear latent variable models, such as PCA and its extensions, are employed to learn reproducible functional connectivity networks present across a cohort of subjects. The activity within each network is subsequently employed as a feature in a linear regression model to predict brain age. The proposed framework is employed on the data from the CamCAN repository and the inferred brain age models are further demonstrated to generalize using data from two open-access repositories: the Human Connectome Project and the ATR Wide-Age-Range.

1 Introduction

The human brain changes during the lifespan of an adult, resulting in robust and reproducible changes in structure and function (Lim et al., 2013; Raz and Rodrigue, 2006). Moreover, there is reason to hypothesize that deviations from the typical brain ageing trajectory may reflect latent neuropathological influences (Cole et al., 2018), serving to motivate further research into developing reliable biomarkers derived from brain imaging data. Such biomarkers could be fundamental in order to better understand and combat age-associated neurodegenerative diseases. To date, early studies have shown success in the context of traumatic brain injury (Cole et al., 2015) and schizophrenia (Koutsouleris et al., 2013).

Due to the significant potential benefits associated with brain-imaging driven biomarkers for age, there have been many statistical models proposed for healthy brain

ageing. These models vary in complexity as well as in the class of neuroimaging data employed. One of the earliest demonstrations was that of Good et al. (2001), who employed voxel-based morphometry to demonstrate the structural changes which occur during healthy ageing. More recently, a wide range of sophisticated machine learning methods have been employed (Franke et al., 2013; Lancaster et al., 2018; Smith et al., 2019). Cole et al. (2015) employed Gaussian process regression to predict the biological age of subjects using structural neuroimaging data, demonstrating that such a model was able to accurately predict brain age. The resulting model was subsequently applied to subjects with traumatic brain injury (TBI), where the associated residuals (difference between predicted and true biological age) were shown to be significantly larger for subjects with TBI as compared with healthy subjects; the associated model consistently predicted subjects with TBI to be *older*, possibly a result of accelerated atrophy. This work was further extended by Cole et al. (2017), who employed convolutional neural networks to obtain improved performance. In related work, Franke et al. (2010) employ kernel regression with an application to the early identification of Alzheimer's disease.

While the vast majority of the literature has employed structural imaging modalities, there are also numerous examples of where functional imaging has been utilized. A pertinent example is Dosenbach et al. (2010), who employ resting-state fMRI together with support vector machines (SVMs) in order to accurately classify subjects as being either children (ages 7-11 years old) or adults (ages 24-30 years old). Furthermore, they observe an overall decrease in network connectivity as subjects mature. In related work, Geerligs et al. (2012) identify ageing-driven changes in functional connectivity, highlighting decreased connectivity within the default mode network and the somatomotor network. Subsequently, Geerligs et al. (2014) categorized the changes in functional connectivity that occur with healthy ageing in terms of various network measures.

More generally, the study of functional connectivity is itself an exciting avenue of modern neuroscientific research which has shown great potential for improving our understanding of the human brain function and architecture (Sporns, 2012). By way of example, changes in functional connectivity have been related to various neuropathologies such as Parkinson's disease (Wu et al., 2009) and Alzheimer's (Damoiseaux et al., 2012) as well as conditions such as Autism (Cherkassky et al., 2006). Recently, the changes in functional connectivity induced by ageing have begun to be studied. Initial studies have reported significant differences in the connectivity between younger and older subjects using resting-state fMRI (Geerligs et al., 2014). Moreover, results appear to suggest there are important changes that occur in the connectivity not just between regions but also at the level of entire networks. However, despite recent advances, a holistic understanding of the relationship between healthy ageing and the associated changes in functional connectivity is still missing.

In this work we seek to build robust models of brain age based on the functional connectivity of individuals. This serves to combine the two prominent avenues of neuroscientific research: brain age prediction and analysis of functional connectivity. In particular, the methods presented in this work have two principal objectives:

1. To demonstrate that measures of functional connectivity can reliably be employed as features in machine learning models of brain age. To this end we build and validate models using three large open-source datasets: the Cambridge Center for Ageing and Neuroscience (CamCAN), the Human Connectome Project (HCP) and the ATR Wide-Age-Range datasets.

2. We further wish to interpret and inspect the proposed models in order to gain further insights into the changes in functional connectivity associated with ageing. This calls for the use of parsimonious and simple predictive models together with features whose relationship with functional conncetivity is clearly understood.

Throughout this paper, we put forward the thesis that for the potential impact of functional connectivity assessment to be met (i.e., in terms of developing powerful biomarkers) the research community needs to develop robust methods for data-analysis which can combine both supervised and unsupervised models of functional connectivity analysis. Instead of tweaking existing statistical methods, it is imperative to develop methods which are intuitive, interpretable, and insightful from a neurophysiological perspective. Such models must utilise as much experimental information as possible in order to investigate the factors which affect functional connectivity.

To further motivate our thesis, one should consider that most experiments to date operate on data from a single laboratory, or class of experiment which limits the generality of any obtained results. Such concerns have been recently recognised, particularly within the context of brain ageing (Geerligs et al., 2015, 2017), and have given rise to multi-laboratory collaborations with data-sharing becoming more common. However, it is still highly unlikely that all subject features (and how these are measured) will be comparable across different experimental environments. Thus while data-sharing has seen much progress, it could be argued that the impact of these endeavours is still to come, and to achieve this, we need to develop methods which can combine information from across disparate, but informative experiments.

To this end we proceed in a two-step framework. First, we seek to learn robust features which summarize properties of functional connectivity across a cohort of subjects in an unsupervised manner. Due to our focus on interpretability, we focus on linear latent variable models, such as principal component analysis (PCA), independent component analysis (ICA) and their generalizations. The benefit of employing latent variable models such as PCA is that we may interpret the latent variables in terms of activity within functional connectivity networks, as proposed by Leonardi et al. (2013) (see also Figure 2 below). Second, once features have been obtained in an unsupervised manner, they are subsequently used to predict brain age using standard linear regression models. We deliberately restrict ourselves to simple linear classifiers as they can be easily interrogated, allowing us to explicitly understand how each feature contributes to the predicted brain age. An overview of our two-stage approach is provided in Figure 1.

The remainder of this manuscript is organized as follows: in Section 2 we first review linear latent variable models and their implications for functional connectivity analysis. We then present our proposed two-step procedure. Experimental results, studying synthetic as well as real resting-state fMRI data, are presented in Section 3.

2 Materials and methods

We focus our analysis on resting-state fMRI time series data which is collected across a cohort of N subjects. For the ith subject, it is assumed we have access to fMRI measurements over p fixed regions of interest, denoted by $X^{(i)} \in \mathbb{R}^p$, as well as the subjects age, $a^{(i)} \in \mathbb{R}_+$. Throughout this work we approximately model the fMRI data for each subject with a stationary multivariate Gaussian distribution, $X^{(i)} \sim \mathcal{N}(0, \Sigma^{(i)})$, where $\Sigma^{(i)}$ denotes the covariance for subject i. Each entry in $\Sigma^{(i)}$ denotes the covariance between any pair of regions, which serves to define a measure of the functional connectivity (Smith, 2012). As such, it follows that $\Sigma^{(i)}$ encodes a functional connectivity network over p regions where edges encode the marginal dependence structure.

The goal of the proposed methods is to learn interpretable and robust models to predict the biological age, $a^{(i)}$, of subjects given information relating only to their functional connectivity. To achieve this, we propose a two-step framework. Our approach first employs linear latent variable models in order to model high-dimensional connectivity matrices using a reduced number of latent variables. We interpret such

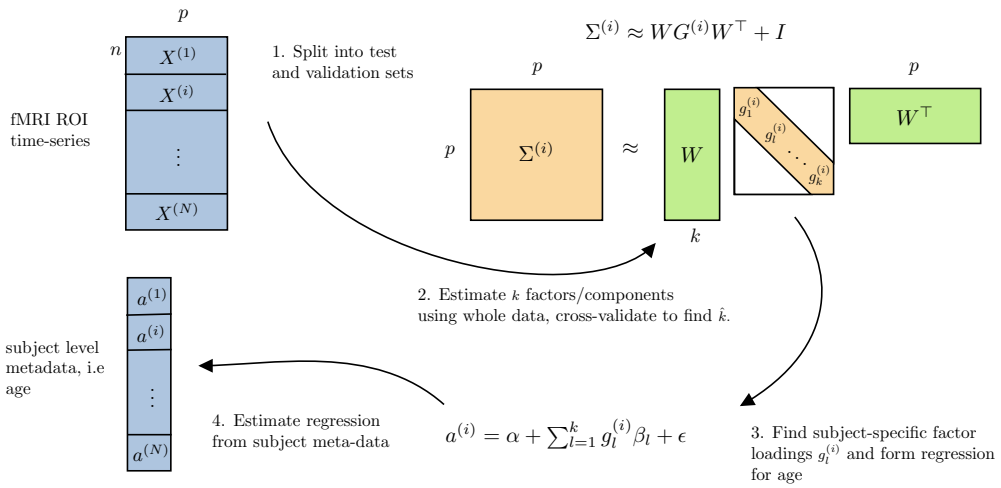

Fig 1. Pipeline for estimating networks, factor loadings, and predictive model for biological brain age. Inferred factors $W \in \mathbb{R}^{p \times k}$ describe networks which are reproducible across the entire population, the subject-specific factor loadings $g_l^{(i)}$ are then used to predict brain age. Once the factor loadings are estimated as above, using one experimental data-set (we use CamCAN data in our experiments), we can then assess how these factors perform for brain age prediction on completely held-out data-sets; we demonstrate how the model generalizes well using HCP and ATR Wide-Age-Range datasets.

variables as corresponding to functional connectivity networks, allowing us to describe patterns in connectivity as being composed of various distinct networks. We note that such a two-step approach has previously been employed in the context of brain age prediction (Franke et al., 2010; Smith et al., 2019). However, as far as we are aware, this is the first work to directly interpret the role of linear latent variable models, such as PCA, as learning the relevant functional networks. This work thereby provides a clear motivation and interpretation for such a two-stage strategy.

In Section 2.1 we discuss the various latent variable models employed, and highlight how introducing assumptions such as non-negativity can help further improve interpretability of results. We also discuss theoretical benefits associated with such assumptions. We then discuss the how the features (i.e., functional networks) inferred by the latent variable models may be used to build linear models for brain age.

2.1 Linear latent variable models for functional connectivity: PCA and its extensions

In this section we outline the linear latent variable models employed in the unsupervised learning stage of the proposed framework. We begin by discussing principal component analysis (PCA), a well-established technique for dimensionality reduction (Jolliffe, 2011). The common derivation for PCA poses it as an optimization problem seeking to learn the linear projection which maximizes explained variance within the projected space (Hotelling, 1933). However, PCA can also be derived as inference under a simple linear latent variable model, which posits that observations $X^{(i)} \in \mathbb{R}^p$ are generated as a linear projection from low-dimensional latent variables, $Z^{(i)} \in \mathbb{R}^k$ (Harman, 1960). When both observations and latent variables are taken to follow a multivariate Gaussian

distributions we obtain the following generative model for observed data:

$$Z^{(i)} \sim \mathcal{N}(0, G^{(i)}) \tag{1}$$

$$X^{(i)}|Z^{(i)} = z^{(i)} \sim \mathcal{N}(Wz^{(i)}, v^{(i)}I) \tag{2}$$

where $G^{(i)} \in \mathbb{R}^{k\times k}$ is a diagonal matrix and $v^{(i)} \in \mathbb{R}_+$ denotes measurement noise. Equations (1) and (2) serve to highlight how PCA can be seen as a low-rank model for the covariance matrix; by marginalizing over latent variables we obtain:

$$\Sigma^{(i)} = WG^{(i)}W^T + v^{(i)}I, \tag{3}$$

implying that the loading matrix, W, captures low-rank covariance structure. Learning the associated loading matrix, W, proceeds via maximizing the log-likelihood over observations across all N subjects:

$$\mathcal{L} = \sum_{i=1}^{N} p \log 2\pi + \log\det\, \Sigma^{(i)} + \text{tr}\left(\Sigma^{(i)^{-1}} K^{(i)}\right), \tag{4}$$

where $\Sigma^{(i)}$ is as defined in equation (3) and $K^{(i)}$ denotes the sample covariance matrix for the ith subject. In the context of PCA, the maximization is performed subject to the constraint that W be orthonormal,

$$\hat{W} = \underset{W:W^TW=I}{\arg\max}\, \{\mathcal{L}\}, \tag{5}$$

and a closed-form solution is obtained via eigendecomposition.

Following Leonardi et al. (2013) it is possible to interpret each column of W as encoding functional networks or "eigenconnectivities". While the loading matrix, W, is shared across all subjects, each diagonal entry of $G^{(i)}$ denotes the extent to which the associated network is expressed in subject i. This allows us to study connectivity as being composed of various distinct networks, resulting in significant benefits from the perspective of interpretability. We can further unpack equation (3) as follows (see also Figure 2 below):

$$\Sigma^{(i)} = \sum_{j=1}^{k} g_j^{(i)} W_j W_j^T + v^{(i)}I, \tag{6}$$

where W_j denotes the jth column of W and we write $g_j^{(i)}$ to denote the jth diagonal entry of the matrix $G^{(i)} \in \mathbb{R}^{k\times k}$. As such, we may interpret each W_j as encoding the jth network and $g_j^{(i)}$ as a measure of activity within the corresponding network in the ith subject.

There exist several extensions to the model described in equations (1) and (2), the prime example being factor analysis which allows the variances in equation (2) to vary across dimensions. Recently, several extensions have been proposed where constraints such as non-negativity are introduced with the goal of improving the interpretability of results (Hirayama et al., 2016; Sigg and Buhmann, 2008; Zass and Shashua, 2007). The motivation behind such methods stems from the fact that interpreting and visualizing PCA-based networks becomes very challenging, particularly in high-dimensions. Challenges arise from the fact that each principal component will correspond to a weighted sum of BOLD activities across all observed regions. As such, it is often difficult to identify which regions are the principal contributors to a certain principal component (and hence functional network) without applying ad-hoc post analysis. Furthermore, it is possible that some entries in the principal components may be negative, which further complicates the interpretation from the perspective of functional connectivity analysis.

The aforementioned issues can be mitigated via the introduction of non-negativity constraints on the loading matrix, W. This ensures that each principal component corresponds only to a weighted *positive* sum of activity over all brain regions. As such, the principal component can be directly interpreted as the contribution of each region to each functional network. Furthermore, the introduction of non-negativity will often yield sparsity in the sense that many of the entries of the principal components will be exactly zero (Sigg and Buhmann, 2008). It follows that such sparsity further facilitates the interpretation of the corresponding networks. From an optimization perspective, the loading matrix is inferred by maximizing the original log-likelihood objective, with the additional non-negativity constraint:

$$\hat{W} = \underset{W:W\geq 0}{\arg\max} \{\mathcal{L}\}. \tag{7}$$

It is important to note that the orthonormality constraint has been dropped in equation (7), making the associated optimization problem less challenging. However, the combination of non-negativity and orthonormality, as enforced in Monti and Hyvärinen (2018), leads to several desirable properties. First, the loading matrix W has at most one non-zero entry per row. This implies that we may interpret the columns of W as encoding membership to k non-overlapping networks or clusters. Another very important benefit of introducing non-negativity and orthonormality constraints is that the matrix W is uniquely defined and identifiable. This is not the case in standard factor analytic models, where W is only identifiable up to an arbitrary rotation (Bishop, 2006; Harman, 1960). Given that throughout this work we will directly interpret the columns of the loading matrix, W, as encoding functional connectivity networks, the lack of identifiability in PCA and factor analysis models is a significant limitation. We refer to the model presented in Monti and Hyvärinen (2018) as Modular Hierarchical Analysis (MHA). The associated optimization problem therefore becomes:

$$\hat{W} = \underset{W:W^TW=I \text{ and } W\geq 0}{\arg\max} \{\mathcal{L}\}. \tag{8}$$

MHA can therefore been seen to address the two important limitations of traditional models such as PCA and factor analysis; first that the presence of negative values in the loading matrix complicates the interpretation of such matrices (addressed via the use of non-negativity constraints) and second is the fact that the latent variables are rotationally invariant (addressed via the further introduction of orthogonality). A further limitation of models such as PCA and factor analysis is that they implicitly assume latent variables must be uncorrelated. In many cases, especially when such models are applied on data relating to a cohort of subjects, such an assumption will not be valid, implying the associated generated models are misspecified. In contrast, MHA is able to identify and recover components even when they are uncorrelated. This is an important theoretical advantage, as MHA continues to enjoy the same identifiability properties even in the presence of correlated latent variables, and practical advantage, as we demonstrate in this work. Finally, we note that in the context of fMRI data, MHA corresponds to an intuitive generative model whereby latent variables capture the activity within each functional network. The optimization of equations (5), (7) and (8) is discussed in Supplement S1. Furthermore, we provide both Python and R code to implement MHA in Supplement S2.

Moreover, we note that model introduced by Hirayama et al. (2016), termed Modular Connectivity Factorization (MCF), shares many similarities with MHA. In fact, both methods introduce non-negativity and orthonormality over the loading matrix, W. The fundamental difference, however, is that MCF is not associated with a

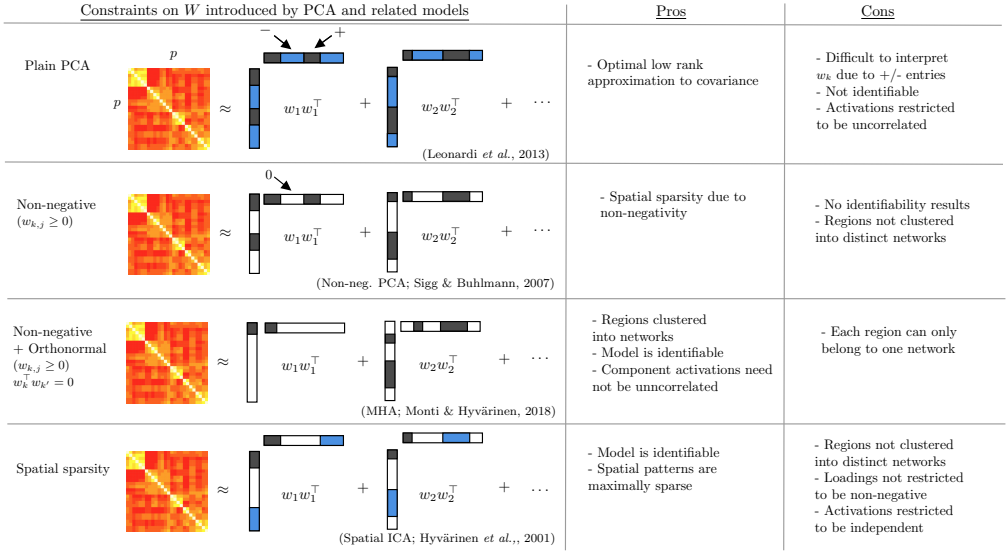

Fig 2. Figure demonstrating the relationship between linear latent variable models, such as PCA and its extensions, to inferred networks. We highlight how introducing various structural constraints on the loading matrix, W, improves interpretability of such models.

linear latent variable model, and instead parameters are inferred as follows:

$$\hat{W} = \underset{W:W^T W=I \text{ and } W\geq 0}{\arg\max} \left\{ \sum_{i=1}^{N} \text{tr}\left(\Sigma^{(i)} K^{(i)}\right)^2 \right\}, \tag{9}$$

where $\Sigma^{(i)}$ is defined as in equation (6) and $K^{(i)}$ is the empirical covariance for the ith subject. A related approach was also proposed by Hyvärinen et al. (2016).

Finally, it is important to note that whilst identifiability can be obtained via the combination of non-negativity and orthonormality, as is the case with the MHA model, it can also be obtained by relaxing the assumed distribution over latent variables, as is the case with independent component analysis (ICA) models. Formally, ICA is also a linear latent variable model, however, latent variables are no longer assumed to follow a Gaussian distribution (Hyvärinen et al., 2001). While the relaxation of the Gaussianity assumption complicates the associated optimization, which must now be solved using gradient descent methods and accounting for the presence of multiple local optima due to the non-convex objective function (Himberg et al., 2004), ICA has been widely employed in the study of functional connectivity (Esposito et al., 2005; van de Ven et al., 2004). Moreover, we note that the "spatial" version of ICA used in fMRI reverses the roles of latent variables and loadings, which means that it is actually looking at the non-Gaussianity or sparsity of what we call here the loadings, corresponding to spatial patterns. Figure 2 provides a visualization of the benefits obtained by introducing each of the aforementioned constraints. In particular, we note that it is the combination of non-negativity together with orthonormality which yields interpretable and identifiable networks. We empirically validate such claims by applying all of the aforementioned models to synthetic and real fMRI datasets below.

2.2 Predicting brain age using functional network activity

The previous section outlined the various flavours of latent variable models which can be employed in order to learn functional networks across a cohort of N subjects. The

aforementioned models allow us to decompose observed functional connectivity patterns as a linear sum of networks encoded by the columns of the loading matrix, W. While the loading matrix is shared across all subjects (indicating the same networks are present across all subjects), the extent to which they contribute to the observed covariance of the ith subject is denoted by the diagonal entries of $G^{(i)}$, as stated in equation (6).

We now consider the task of predicting the biological brain age, $a^{(i)}$, using inferred functional connectivity networks as features. In the interest of interpretability we limit ourselves to linear regression models of the form:

$$a^{(i)} = \sum_{j=1}^{k} \beta_j g_j^{(i)} + \epsilon^{(i)}. \tag{10}$$

Recall that $g_j^{(i)}$ corresponds to the jth diagonal entry of the matrix $G^{(i)}$. As such, the proposed models will essentially seek to predict the biological age of subjects by considering activity within each inferred functional network. In the case of the ith subject, the observed activity in network j is quantified by $g_j^{(i)} \in \mathbb{R}_+$. In practice, we will seek to quantify the activity of various functional networks on unseen subjects, defined to be subjects whose data was not employed to estimate loading matrix, W. We note that due to the orthonormality of W, together with equation (6), we may estimate $g_j^{(i)}$ for data from unseen subjects, denoted by i^*, as follows:

$$\hat{g}_j^{(i^*)} = W_j^T \hat{\Sigma}^{(i^*)} W_j - v^{(i^*)}. \tag{11}$$

We note that equation (11) requires the observation noise, $v^{(i^*)}$. This is not a concern for all subjects whose data is employed during the unsupervised learning of the latent variables, as parameters $v^{(i)}$ are inferred alongside loading matrix, W. However, the primary goal of this work is to build predictive models which can generalize to unseen subjects. In this context, an estimate of the observation noise, $v^{(i^*)}$, can be obtained as follows:

$$\hat{v}^{(i^*)} = \text{tr } \hat{\Sigma}^{(i^*)} - W^T \hat{\Sigma}^{(i^*)} W. \tag{12}$$

Although the class of models considered in equation (10) may be considered amongst the simplest supervised regression models, they yield several important benefits when seeking to understand both the estimated parameters as well as the contribution of each of the features. In particular, each β_j corresponds to the regression coefficient summarizing the (linear) relationship between the activity of the jth network and biological age, conditional on all remaining networks. As such, if certain regression coefficients are deemed to be insignificant, we may conclude that the associated network is invariant during healthy ageing.

2.3 Hyper-parameter selection

The proposed two-stage estimation framework requires the input of only one hyper-parameter: the dimensionality of latent variables k. In the context of PCA and factor analysis, this hyper-parameter directly corresponds to the number of principal components or factors inferred, and a wide literature exists for tuning such a parameter (Jolliffe, 2011). One of the advantages of the latent variable models presented in Section 2.1 is that they each correspond to probabilistic models whose likelihood can be directly evaluated. As such, a logical choice to tuning hyper-parameter k is to directly maximize the log-likelihood over held out data.

In order to effectively perform hyper-parameter tuning as well as quantify the generalization performance of the proposed method, data was split into training, validation and test datasets as follows:

- First, a subset of subjects were held out as test data. As such, we obtain two datasets:
$$\left\{X_{1:n}^{(i)}, a^{(i)}\right\}_{i \in S_{train}} \quad \text{and} \quad \left\{X_{1:n}^{(i)}, a^{(i)}\right\}_{i \in S_{test}}$$
where $S_{train}, S_{test} \subset \{1, \ldots, N\}$ denote the non-overlapping sets of training and test subjects respectively. Recall N is the number of subjects present and we write $X_{1:n}^{(i)}$ to denote the n observations available for the ith subject.

- Training data is further split into training and validation datasets on a subject-by-subject basis.

Splitting the data in this manner allows for effective hyper-parameter tuning, using training and validation datasets, as well as for generalization performance to be measured using test dataset which corresponds to unseen subjects.

2.4 Experimental data

The data employed in this manuscript corresponds to resting-state fMRI data taken from three distinct open-access repositories. There were small variations in the resting-state functional MR image acquisition for each of the repositories considered: CamCAN (Taylor et al., 2015), Human Connectome Project (Van Essen et al., 2013), and the ATR Wide Age Range (Ogawa et al., 2018). The pre-processing employed on each dataset was as follows:

- CamCAN: This dataset was pre-processed by us. Data was motion corrected, spatially smoothed with a 5mm FWHM Gaussian kernel, registered into MNI152 standard space using FLIRT (Smith et al., 2004) via a skull-stripped high-resolution T1 image and resampled to 4x4x4mm voxel sizes. Each high resolution T1 image was segmented into grey and white matter and cerebrospinal fluid using SPM Dartel (Ashburner, 2009). Mean timecourses for cerebrospinal fluid and white matter as well as 6 motion parameters were linearly filtered from each voxel to reduce non-neural noise.

- HCP: We used the pre-processed resting-state fMRI data from a random subset of healthy participants[1]. Notably, the pipeline involved FIX ICA-based noise reduction process (Salimi-Khorshidi et al., 2014), to remove individual sources of physiological, non-physiological and motion related noise.

- ATR: We used the preprocessed data[2]. The pre-processing pipeline notably included regressing out the global grey matter signal as well as signals from cerebrospinal fluid and white matter, to remove sources of spurious variation.

All three pre-processed fMRI datasets were subsequently processed as follows: a cortical parcellation based on resting-state functional connectivity analyses (Power et al., 2011) was used to define 264 distinct 10mm diameter regions of interest (ROIs). The fMRI time course averaging across all voxels within each ROI was extracted. These 264 average time courses were then used in subsequent analyses.

[1]Full details of the pre-processing pipeline can be found at `https://www.humanconnectome.org/study/hcp-young-adult/document/extensively-processed-fmri-data-documentation`

[2]Full details are provided here `https://bicr-resource.atr.jp/var/www/webapp/bicrresource/bicrresource/staticfiles/pdf/Methods.pdf`

3 Results

In this section we present a range of experimental results involving both synthetic and real resting-state fMRI datasets. Throughout this section, we contrast the performance of the various linear latent variable models presented in Section 2.1. In particular, we study the performance across the following methods[3]: factor analysis (FA), PCA, non-negative PCA (Sigg and Buhmann, 2008), MCF (Hirayama et al., 2016) and MHA (Monti and Hyvärinen, 2018) as well as ICA. In the case of ICA, we first employ PCA as a dimensionality reduction before employing the FastICA algorithm proposed by Hyvärinen (1999).

We first present results using synthetic data in Section 3.1. These simulation experiments serve as a numerical validation of the proposed two-stage procedure. Experiments relating to brain age prediction from resting-state fMRI data are subsequently presented in Section 3.2.

3.1 Synthetic data experiments

In this section we evaluate the performance of the proposed two-stage estimation framework using synthetic data. To this end, we generate artificial data whose properties approximately match those which are frequently reported in fMRI studies. The objective is then to quantify which of the linear latent variable models presented in Section 2.1 are able to both robustly recover the associated loading matrix, W, as well as learn the relevant factors which serve as accurate predictors of brain age on unseen subjects. Synthetic data was then generated in order to satisfy equations (1-2) and (10). This is achieved as follows:

- First, we randomly generated a factor loading matrix, $W \in \mathbb{R}^{p \times k}$, which satisfied the constraints of both non-negativity and orthonormality. The reason for introducing both constraints is that we will seek to quantify how reliably each latent variable model can recover W, and it is therefore imperative to ensure we generate W from an identifiable model (see discussion in Section 2.1). In order to achieve this a dense matrix, W, was sampled with each entry following a uniform distribution over the interval $[0, 1]$. Subsequently, for each row only the entry with the largest value was retained with all other entries set to zero. Finally, the norm of each column was set to one.

- Second, the factor loadings for the ith subject, $g^{(i)} \in \mathbb{R}^k$, were randomly generated as follows:

$$g_j^{(i)} \sim \mathcal{N}(2.5, 1.0), \quad \text{for } j = 1, \ldots, k$$

  with all negative samples being discarded.

- The regression coefficients, $\beta \in \mathbb{R}^k$, were drawn uniformly at random from the interval [0,10].

- Finally, we are able to randomly generate observations and ages for each subject as follows:

$$X^{(i)} \sim \mathcal{N}(0, WG^{(i)}W^T + v^{(i)}), \tag{13}$$

$$a^{(i)} \sim \mathcal{N}(\beta^T g^{(i)}, \epsilon). \tag{14}$$

  Recall that $G^{(i)} \in \mathbb{R}^{k \times k}$ is a diagonal matrix consisting of entries $g_j^{(i)}$.

[3]The implementations available in `Scikit Learn` were employed for Factor Analysis, PCA and ICA (Pedregosa et al., 2011).

We note that the choices for sampling distributions of both the factor loadings, $g^{(i)}$, as well as the regression coefficients, β, are necessarily somewhat heuristic. However, care was taken to ensure the implied distributions over subject ages approximately matched the empirical distributions observed within the CamCAN repository.

We note that throughout experiments we consider the performance of each method whilst varying two distinct factors: the number of observations per subject, n, and the number of training subjects, N. Furthermore, throughout simulations we fix the dimensionality of observations to be $p = 50$ and the number latent factors to be $k = 5$.

Given artificial data generated as described above, we look to quantify the performance of each of the linear latent variable models using the following two metrics:

1. Accurate recovery of the loading matrix, W. This is quantified in terms of the squared error between the true loading matrix and the estimated loading matrix.

2. Accurate brain age prediction over unseen subjects. In line with other literature, this is quantified in terms of the mean absolute error between true and predicted brain ages (Franke et al., 2010; Lancaster et al., 2018).

3.1.1 Synthetic data results

We begin by considering the performance of each linear latent variable model as the number of observations per subject, n, increases for a fixed number of training subjects, $N = 25$. The results are presented in Figure 3. We note that both in terms of recovery of the loading matrix, W, as well as in terms predicting the ages over unseen subjects, the introduction of regularity constraints, be they in the form of non-negativity, orthonormality or non-Gaussianity or sparsity (as in ICA), leads to improvements.

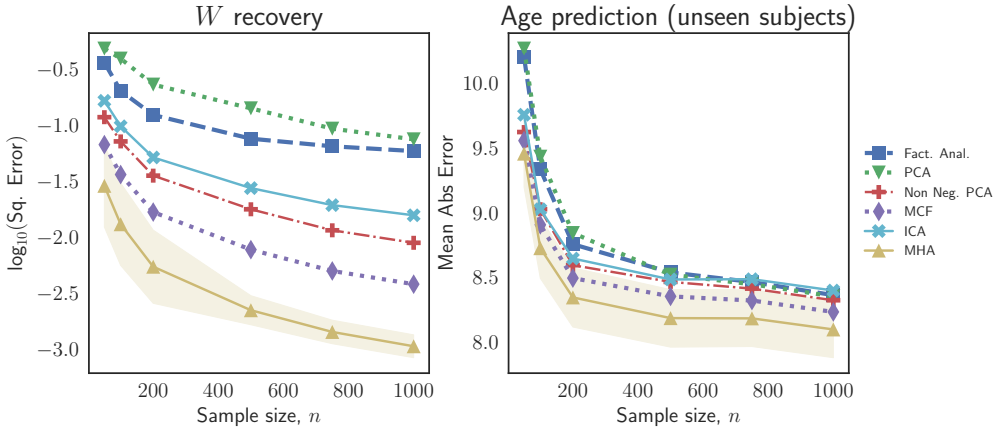

Fig 3. Simulation results for recovery of the true loading matrix (left panel) and prediction of brain age for unseen subjects (right panel) as the number of observations per subject, n, increases. We note that the introduction of regularity constraints (e.g., non-negativity or orthonormality) on the loading matrix leads to improvement in performance.

We also study the performance of the various latent variable models when the number of training subjects, N, increases and the number of observations is fixed at $n = 100$ per subject. These results are presented in Figure 4. In terms of recovery of the loading matrix, W, we again observe that introducing regularity constraints leads to significant improvements. In terms of predictions over unseen subjects (as shown in the

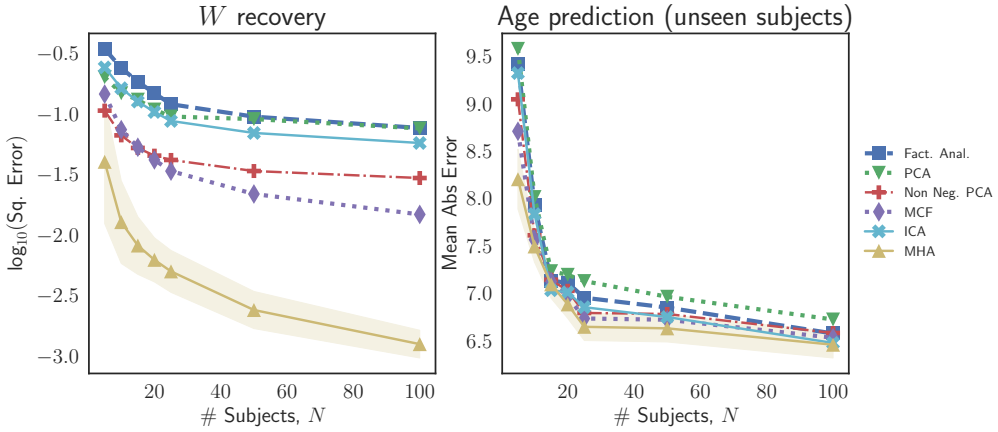

Fig 4. Simulation results for recovery of the true loading matrix (left panel) and prediction of brain age for unseen subjects (right panel) as the number of training subjects, N, increases. We note that the introduction of regularity constraints (e.g., non-negativity or orthonormality) on the loading matrix leads to improvement in performance.

right panel of Figure 4), the improvements due to the introduction of regularity conditions begin to fade as the number of training subjects increases. In particular, beyond a certain number of training subjects (approximately 25 in the case of these experiments), the improvement in out-of-sample predictions begins to plateau.

3.2 Resting-state fMRI data experiments

While the previous section presented results relating to synthetic data, here we present experimental results where the proposed two-step procedure is applied to three open-source resting-state fMRI datasets. The datasets considered correspond to the Cambridge Center for Ageing and Neuroscience (CamCAN) repository, the Human Connectome Project (HCP) repository, and the ATR Wide-Age-Range repository. The purpose of employing three distinct datasets is to effectively measure the generalization performance of the proposed approach on unseen data. As such, data from the HCP and Wide-Age-Range repositories was not employed during any of the model training and instead used exclusively as unseen test data. It is important to note that in addition to significant inter-subject variability (Kelly et al., 2012), fMRI data also suffers from the presence of several other well-documented issues such as variable scanner performance or noise (Bennett and Miller, 2010; Friedman et al., 2006; Poldrack et al., 2011). As such, validating the performance of the proposed brain age prediction models in this way will provide a more realistic measure of their generalization performance.

3.2.1 CamCAN repository results

Resting-state fMRI data was collected from a total of 647 subjects from the CamCAN repository. Subject ages ranged from 18 to 88 years of age (average age of 54.31±18.56, 318 males and 329 females). The CamCAN dataset was employed as the principal dataset in the proposed two-step procedure, implying that it was employed to learn both the functional network structure in the unsupervised learning stage and the linear regression models in the supervised learning stage. As such, the data was split into

training, validation and test subsets as described in Section 2.3.

Step 1: unsupervised functional network inference

The first stage of the proposed framework involves the estimation of reproducible functional connectivity networks via the use of the various linear latent variable models discussed in Section 2.1. The number of functional networks inferred corresponds directly to the dimensionality of latent variables, which is determined by hyper-parameter k. As each linear latent variable model can be interpreted as a probabilistic model, we select hyper-parameter k by maximizing the log-likelihood over the validation dataset. This resulted in the choice of $k = 5$ when the loading matrix was restricted to be both non-negative and orthonormal, as proposed by Hirayama et al. (2016) and Monti and Hyvärinen (2018). While it is possible that the choice of hyper-parameter may vary across distinct latent variable models (e.g., for PCA or factor analysis), we choose to keep the choice of k fixed across all models as this facilitates model comparison and interpretation of results.

The left panel of Figure 5 visualizes the results when the MHA linear latent variable model was employed[4]. We note that, as discussed in Section 2.1, the MHA linear latent variable model effectively clusters regions into sub-networks via the introduction of non-negativity and orthonormality constraints. As such, each plot in the left panel of Figure 5 visualizes spatially remote brain regions which have been clustered together, indicating that these regions share strong positive correlations. We note that these correlations (i.e., edges in a network) are omitted for clarity in Figure 5. The results demonstrate that the inferred networks are spatially homogeneous and symmetric across both hemispheres. Furthermore, many of the inferred networks correspond to widely reported networks and regions: network 1 captures the default model network (DMN) and network 2 overlaps with the salience network, while networks 3 and 4 correspond to a higher-level visual network and the somatomotor network respectively. For comparison, we include equivalent plots for all other latent variable models considered in visualized in Figure 13, presented in the Supplementary Material. We note that alternative methods, such as PCA, which did not enforce the combination of both non-negativity and orthonormality, yielded results which were visibly less clustered and more difficult to interpret.

The right panel of Figure 5 visualizes the correlation between the activity of each network (as defined in equation (11)) with the age of each subject. For networks 1-3 we observe a significant negative correlation between the activity and age, suggesting that ageing induces a drop in activity of such networks. These results are in line with related research on ageing induced differences in functional connectivity. In particular, the decrease in activity of the DMN (network 1), has been widely reported (Geerligs et al., 2015; Grady et al., 2016; Liem et al., 2019).

Step 2: supervised training of brain age prediction models

Recall that the overall objective of the proposed framework was build interpretable models of biological brain age. To this end, the features recovered from linear latent variable models where employed as features in a linear regression framework to predict the brain age of each subject. In particular, the five distinct the linear latent variable models detailed in Section 2.1 where employed to learn reproducible sub-networks parameterized by a loading matrix, $W \in \mathbb{R}^{p \times k}$. The activity within each functional network, defined as in equation (11), was subsequently employed as features to predict biological age using linear regression.

[4]Figures produced using the `plot_glass_brain` function from the `nilearn` python module (Abraham et al., 2014).

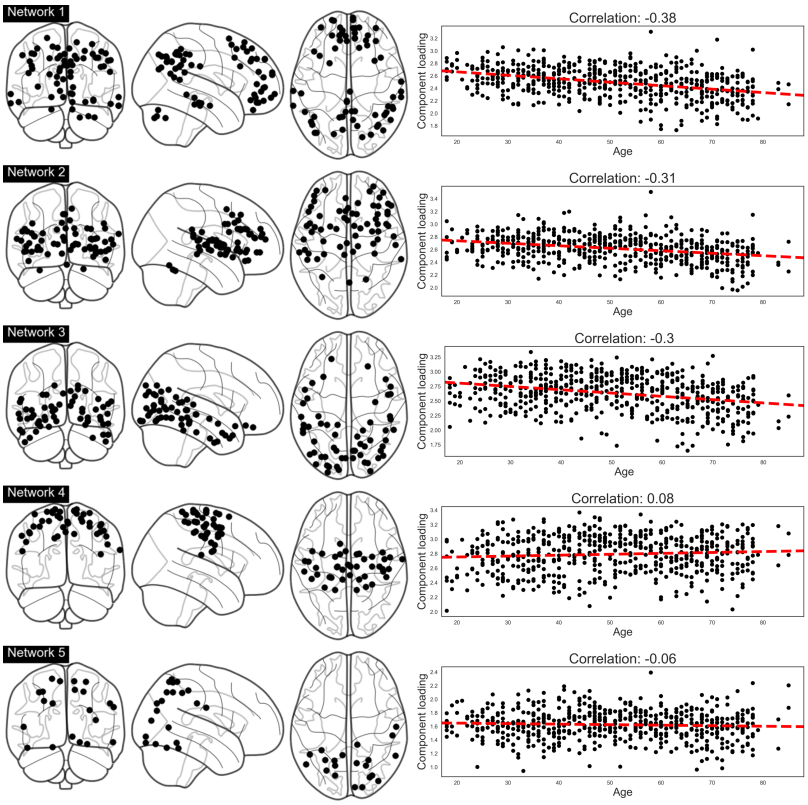

Fig 5. Left panel: inferred networks as recovered when non-negativity and orthonormality constraints are introduced over the loading matrix, W. Networks are spatially consistent and symmetric. Right panel: visualization of network activities against subject age demonstrating (mostly negative) linear trends with healthy aging.

We note that the CamCAN repository, as well as HCP and ATR repositories, each contained over a hundred subjects each. This is in contrast to typical fMRI studies, where the sample size is often in the range of 20 to 30 subjects (Cremers et al., 2017; Poldrack et al., 2011). Furthermore, recall that the goal of experiments presented are to quantify performance on unseen resting-state fMRI data with a view to providing an indication of how each of the linear latent variable models employed would perform in a typical fMRI study. As such, throughout the remainder of this section we report the performance, in terms of mean absolute error, over random subsets of 30 subjects from each repository. This corresponds to a form of bootstrapping, where we average results over a random sample of possible *cohorts*. In practice, we report results over 1000 random subsets of 30 subjects for each of the three repositories considered.

Figure 6 visualizes the mean absolute error on unseen test data for various choices of $k \in \{2, \ldots, 10\}$. We note that the combination of linear regression with the use of non-negativity and orthonormality constraints, as advocated by both the MCF and MHA models, leads to competitive performance over a range of choices of k. In particular, such algorithms out-perform both non-negative PCA and PCA, suggesting that the introduction of such constraints serves to improve the predictive properties of the model. Moreover, we note that Figure 6 indicates the presence of a bias-variance trade-off that is often encountered in supervised learning whereby performance on unseen test data begins to deteriorate as the number of parameters (in our case k) increases beyond a certain value.

As mentioned previously, the choice of $k = 5$ was selected in by maximizing

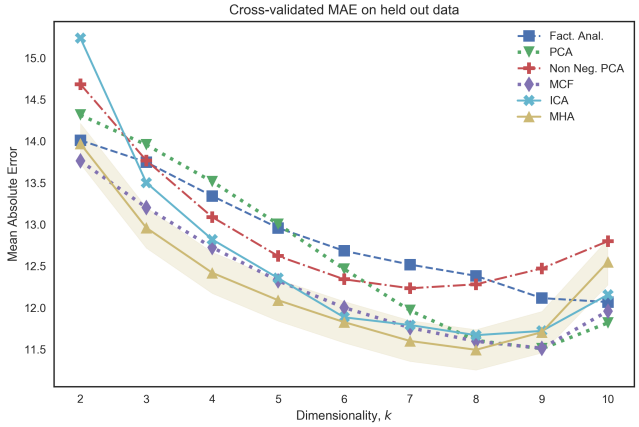

Fig 6. Mean absolute error (MAE) performance for a varying number of networks, as determined by k (x-axis), on unseen test data from CamCAN. We note that the combination of non-negativity and othonormality (MHA and MCF) yields competitive results across a wide range of k.

log-likelihood over a validation dataset (i.e., in an entirely unsupervised manner - data regarding subject ages was not considered). Figure 7 visualizes the performance on the unseen test dataset for the specific choice of $k = 5$, for all possible choices of linear latent variable models. The results indicate that as additional constraints are introduced to the loading matrix, the generalization capabilities of the models also improve. As such, MCF and MHA, which introduce the most stringent constraints corresponding to *both* non-negativity and orthonormality, obtain the best generalization performance. We also note that ICA is also competitive. Moreover, non-negative PCA, which relaxes the requirement for orthonormality, is the next most competitive latent variable model. Finally, PCA and factor analysis, which relax all the aforementioned constraints, obtain the worst generalization performance.

3.2.2 Transfer onto HCP and ATR Wide-Age-Range repositories

The results of Section 3.2.1 provide a measure of performance, in terms mean absolute error in predicted brain age, within a large-scale resting-state fMRI dataset. However, it is widely accepted that in addition subject-specific noise, there are several other

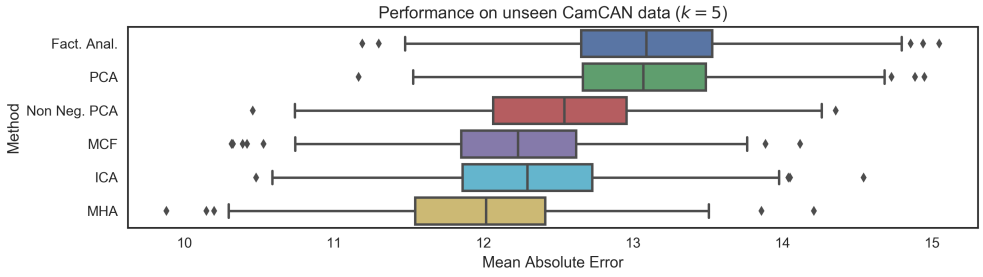

Fig 7. Mean absolute error (MAE) performance on unseen testing data from CamCAN repository when the dimensionality of latent variables is fixed to $k = 5$ (implying we infer 5 networks). We note that as regularity constraints are introduced, in particular non-negativity and orthonormality, predictive performance improves.

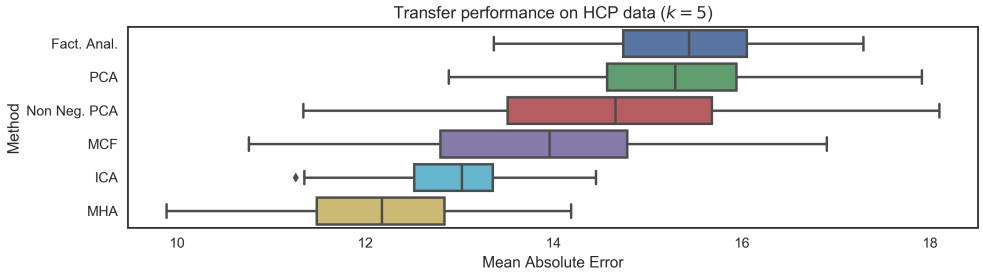

Fig 8. Mean absolute error (MAE) performance on unseen data from HCP repository. Results are broadly consistent with performance on the CamCAN data, indicating good generalization. We note that the introduction of non-negativity or orthogonality constraints leads to improved generalization. The number of functional networks was $k = 5$.

significant contributors to noise in fMRI data: these include issues related to scanner noise and frequency of acquisition of images (Bennett and Miller, 2010; Friedman et al., 2006; Poldrack et al., 2011). As a result, in order to thoroughly verify the generalization performance of the proposed methods, we employ resting-state fMRI data from the HCP and ATR Wide-Age-Range repositories. We note that data from the aforementioned repositories was employed only for testing purposes, as such it was not employed to learn the network structure across subjects, nor to tune the parameters of the linear regression models. For a summary of the characteristics of HCP and ATR Wide-Age-Range datasets see Figure 12 and Table 1 in the Supplementary Material.

Prediction of biological age on both the HCP and ATR Wide-Age-Range repositories was performed as follows: First, the loading matrix, $\hat{W}$ was employed to obtain estimated activity within each network, as detailed in equations (11) and (12). Subsequently, predictions of biological age were obtained using equation (10). At each stage both $\hat{W}$ and $\hat{\beta}$ are the parameters inferred using the CamCAN dataset (i.e., there was no fine-tuning of parameters). As a result, performance on both HCP and ATR Wide-Age-Range datasets provide a robust measure of generalization performance to entirely unseen data.

Results on the HCP data are provided in Figure 8. As expected, the mean absolute errors are larger for each of the distinct latent variable models when compared to the results of on the CamCAN dataset (Figure 7), which will be partially the result of varying scanner noise and image acquisition properties. Importantly we note that, as with the CamCAN dataset, there once again a relationship between the introduction of additional constraints (in the form of non-negativity, orthonormality or non-Gaussianity) and generalization performance. As before, methods such as PCA and factor analysis which do not introduce any constraints had the weakest performance as well as the largest drop in performance.

The HCP results presented above serve to partially validate the predictive models trained using the CamCAN dataset. However, one significant limitation of the HCP dataset is that subject ages only range from 22 to 37 years of age. This is particularly relevant in the context of brain age biomarkers, as many neurodegenerative diseases of interest will be associated with advanced ages. As a result, we further validated the generalization capabilities of the proposed brain age prediction models on the ATR Wide-Age-Range dataset, which had subjects ranging from 20 to 70 years of age. Results, presented in Figure 9 are consistent with results on the CamCAN and HCP datasets, again indicating that the introduction of constraints non-negativity and orthonormality constraints improves generalization performance.

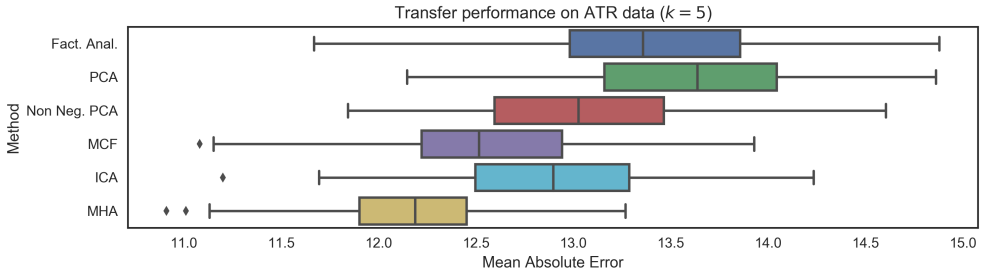

Fig 9. Mean absolute error (MAE) performance on unseen data from ATR Wide-Age-Range repository. Results are broadly consistent with performance on the CamCAN data, indicating good generalization. Further, as with the HCP data, we note that the introduction of non-negativity or orthogonality constraints leads to improved generalization. The number of functional networks considered was $k = 5$.

3.3 Extension to non-independent latent variable models

The results presented above employ linear latent variable models where the inferred latents are assumed to be independent. This is clearly stated in the generative model considered in equation (1) where the covariance of latent variables, $G^{(i)}$, is assumed to be diagonal[5]. However, such an assumption will often fail in practice, implying that the empirical covariance structure over latent variables will not be diagonal. In this section we seek to exploit this by directly introducing the off-diagonal entries of the latent variable covariances, $G^{(i)}$, as features in our linear regression models for biological age. As such, whilst equation (10) considered a linear model where only the diagonal entries

[5]Note that in the case of PCA, factor analysis and MHA, since latent variables are assumed to be multivariate Gaussian, the fact the covariance is diagonal implies the latent variables are independent.

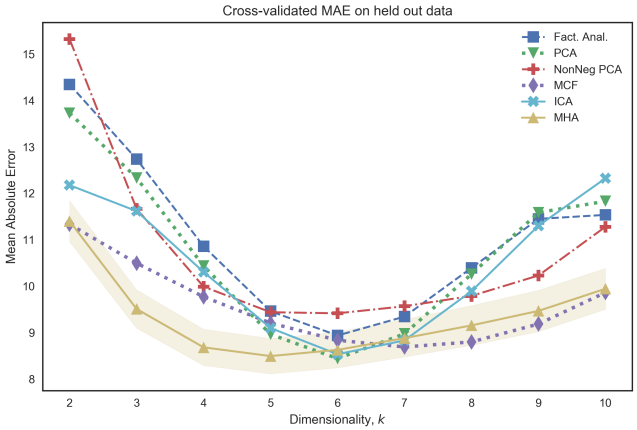

Fig 10. Mean absolute error (MAE) performance for a varying number of networks, as determined by k (x-axis), on unseen test data from CamCAN when latent variables are no longer assumed to have an isotropic covariance structure and the full vectorized covariance is employed as features in the linear regression models. We note that MHA is able to directly accommodate such a scenario and hence is competitive for all choices of latent variable dimension, k.

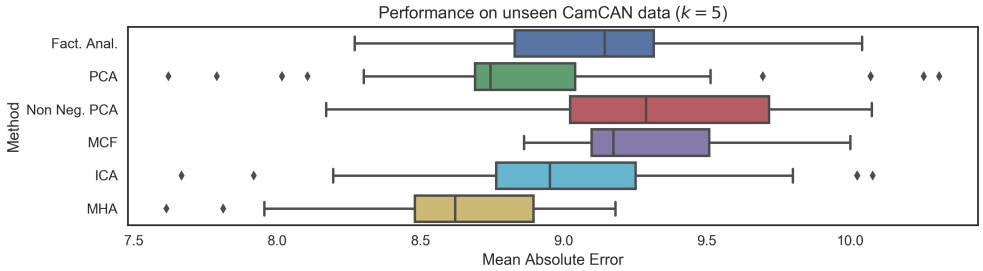

Fig 11. Mean absolute error (MAE) performance on unseen testing data from CamCAN repository when the dimensionality of latent variables is fixed to $k = 5$ (implying we infer 5 networks). Note that latent variables are no longer assumed to have an isotropic covariance structure and the full vectorized covariance is employed as features in the linear regression models.

of each $G^{(i)}$ were employed to predict biological ages of each subject, we now consider linear regression models of the following form:

$$a^{(i)} = \sum_{j=1}^{k} \sum_{l \geq j} \beta_{jl} \; {g_{jl}}^{(i)} + \epsilon^{(i)}. \tag{15}$$

Note that in equation (15) we employ the full upper triangular entries of the covariance matrix as features. This is equivalent to vectorizing the covariance matrix and removing duplicate entries due to symmetry. As such, whilst k features were employed in equation (10), we now consider a linear models with $\binom{k}{2}$ features; many of which will seek to predict the biological age of individuals based on the off diagonal entries of each $G^{(i)}$. It is important to note that the model presented in equation (10) is a special case of equation (15).

As in Section 3.2, we proceed in a two-stage approach whereby we first estimate the loading matrices for the various linear latent variable models employed and subsequently train linear regression models using the full vectorized covariance matrix as features.

Figure 10 visualizes the MAE error on unseen test data as a function of the dimensionality of latent variables, k. We note that for all choices of k the reported errors are smaller than those reported in Figure 6. This provides empirical evidence that the off-diagonal entries of the latent variable covariances are discriminative features for brain age prediction, and therefore can be seen as evidence that models which assume diagonal covariance structure over latents are misspecified. Figure 11 provides further visualizations in the case where $k = 5$. We note that the MHA model performs competitively, this is to be expected as this model directly accommodates the possibility of non-independent latent variables (Monti and Hyvärinen, 2018). Moreover, we note that MHA performs particularly well when the number of networks is small (when dimension of latent variables, k, is less than or equal to 5), which is useful when we wish to prioritize the interpretability of results. Finally, the performance of various methods, as depicted in Figure 10, shows similar trends as in Figure 6; there is once again a bias-variance trade-off associated with the choice of k and the introduction of non-negativity or non-Gaussianity constraints (as in MCF or ICA) leads to improved generalization performance. Finally, whilst Figure 10 only shows generalization performance to unseen subjects from the CamCAN cohort, we also present results for generalization performance to brain age prediction on the HCP and ATR Wide-Age-Range datasets in Figures 14 and 15 of the Supplementary Material.

4 Conclusion

It is widely accepted that ageing has pronounced effects on the functional architecture of the human brain (Geerligs et al., 2014; Smith et al., 2019). In the current study we have presented and validated a two-stage framework through which to train interpretable and robust models of biological brain age based on functional connectivity. In particular, the proposed framework first employs linear latent variable models to uncover reproducible networks which are present throughout a cohort of subjects. A variety of such latent variable models are considered many of which extend PCA by introducing constraints such as non-negativity over the loading matrix. Our experiments suggest that whilst PCA is a natural candidate for dimensionality reduction, and can be interpreted as recovering latent *eigenconnectivities*, the introduction of constraints such as non-negativity can serve to greatly improve both interpretability and predictive performance. While ICA improves on PCA by introducing spatial sparsity, we found that MHA as well as MCF lead to better results, especially in the case of a small number of networks. Reasons for this improvement include using a combination of non-negativity and orthogonality that leads to disjoint networks, as well as explicit modelling of connectivity between the components.

Given inferred functional networks and their activations we train linear predictive models of biological brain age where in the interest of interpretability we deliberately restrict ourselves to linear models. This allows us to directly interrogate the effects of each functional network on the predicted brain age (as shown in Figure 5). In line with other results in the literature, we find a decrease in activation in the default mode network, salience network and higher-level visual network as biological age increases.

The proposed two-stage framework is first validated on the data from the CamCAN repository and subsequently further applied to two further open-access repositories: the HCP and ATR Wide-Age-Range repositories. The use of data from two additional repositories serves to provide a clear empirical indication of the generalization capabilities of the proposed approach. This is especially relevant in the context of fMRI data, where artefacts such as scanner noise can often cause significant challenges (Poldrack et al., 2011).

We note that the brain age prediction errors presented in this work are not competitive with alternative methods which are based on alternative imaging modalities, such as structural imaging data (Cole and Franke, 2017; Cole et al., 2017). This is to be expected for two reasons. First, the imaging modality employed in this work, resting-state fMRI data, is both noiser and likely to be less age-indicative than structural measures. Second, in this work we deliberately restrict ourselves to building simple yet interpretable models of brain age. As such, we restrict ourselves to consider only linear classifiers as these allow for clear model interpretation and interrogation, while noting that the use of more expressive models (e.g., nonlinear models) in the second stage should naturally lead to improved performance.

Furthermore, it is important to note that whilst this work demonstrates the feasibility of functional connectivity driven models of biological brain age, all subjects included in these studies were healthy. As such, whilst such models could eventually be employed to develop biomarkers, further experimentation and validation will be required in future. Moreover, an avenue for further research would be to consider performing classification instead of regression in the second stage of the proposed method. Whilst a natural task would be to discriminate between healthy controls and subjects with some neuropathology, such an approach could also be employed in the context of task-based fMRI as well as to study changes in functional connectivity induced by various distinct tasks (Zippo et al., 2019a) or neuropathologies (Lorenz et al., 2018; Zippo et al., 2019b). In particular, task-based fMRI has been widely reported as displaying non-stationary functional connectivity structure (Calhoun et al., 2014; Monti et al., 2014, 2017a,b). As

such, seeking to discriminate between various cognitive tasks, for example as considered by Chung et al. (2016); Lorenz et al. (2019); Monti et al. (2015, 2017c), could be an exciting future application. Moreover, while in this work we have considered linear latent variable models such as PCA, future work could consider alternative latent variable modes such as latent position graphs (Athreya et al., 2017) and causal models (Khemakhem et al., 2019; Monti et al., 2019; Sasaki et al., 2019).

Supporting information

S1 Appendix. Technical details of the MHA algorithm. In this appendix, we give further details of the block-coordinate descent algorithm which we implement to update the model parameters. In practice, we solve the constrained optimisation (8) via the use of projections onto the non-negative quadrant (non-negativity) and Lagrange multipliers. More specifically, we use the objective function:

$$\tilde{\mathcal{L}} = \mathcal{L} + \frac{\delta}{2}||W^T W - I_k||_2^2 + \text{tr}(\Gamma^T(W^T W - I_k)), \tag{16}$$

where $\Gamma \in \mathbb{R}^{k\times k}$ and δ are Lagrange multipliers enforcing the orthonormality constraints.

We employ gradient descent approach to update the estimate of W. To this end, we follow Monti and Hyvärinen (2018) and introduce a gradient step size η and project onto the non-negative orthant at each iteration (this ensures that the positivity constraint is maintained). The update takes the form

$$W \leftarrow \mathcal{P}^+\left(W - \eta\left(\frac{\partial \mathcal{L}}{\partial W} + \delta(WW^\top W - W) + W\Gamma)\right)\right), \tag{17}$$

where $\mathcal{P}^+ = \max(0, x)$ denotes the projection onto the non-negative orthant and η is a stepsize parameter. The update for the Lagrange multipliers Γ is given by (Bertsekas, 2014):

$$\Gamma \leftarrow \Gamma + \delta(W^\top W - I).$$

In the case of the loading matrix, the gradient update is defined as:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial W} &= \sum_{i=1}^{N} \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial W} \\ &= \sum_{i=1}^{N} \left(-{\Sigma^{(i)}}^{-1} + {\Sigma^{(i)}}^{-1} S^{(i)} {\Sigma^{(i)}}^{-1}\right) W G^{(i)} ,\end{aligned} \tag{18}$$

where we note that via the Sherman-Woodbury identity and using the form of the covariance (3), we can write ${\Sigma^{(i)}}^{-1}$ as follows:

$${\Sigma^{(i)}}^{-1} = (v^{(i)} I)^{-1} - (v^{(i)} I)^{-1} W (G^{(i)^{-1}} + W^\top v^{(i)} I W)^{-1} W^\top (v^{(i)} I)^{-1}. \tag{19}$$

For the diagonal matrix of eigenvalues, $G^{(i)}$, we can update each matrix independently as follows:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial G^{(i)}} &= \frac{\partial \mathcal{L}}{\partial \Sigma^{(i)}} \frac{\partial \Sigma^{(i)}}{\partial G^{(i)}} \\ &= \sum_{i=1}^{N} \left(-{\Sigma^{(i)}}^{-1} + {\Sigma^{(i)}}^{-1} K^{(i)} {\Sigma^{(i)}}^{-1}\right) .\end{aligned} \tag{20}$$

S2 Code. Python and R implementations of the MHA algorithm.

- Python: `https://github.com/piomonti/MHA`
- R: `http://www.gatsby.ucl.ac.uk/~ricardom/FactorCovariance_ScoreMatch_PenalizeLagrange.R`

S3 Fig. Age distributions of subjects across repositories.

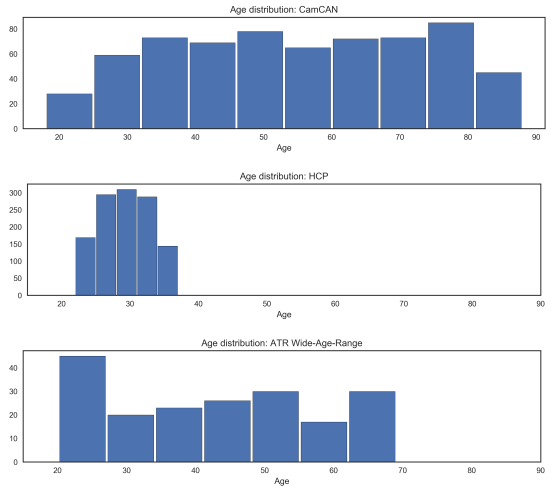

Fig 12. Histogram visualizing age distribution for each of the repositories employed. We note that the CamCAN dataset has the widest range of all repositories considered, validating its use as a the primary dataset in our study.

Dataset	# subjects	Age range
CamCAN	647	18—88
HCP	80	20—36
ATR	191	20—70

Table 1. Table detailing number of subjects studied in each of the three datasets considered. In the case of the HCP datasets, 80 subjects were randomly selected out of all possible subjects.

S4 Fig. Functional connectivity networks inferred by PCA and alternative models

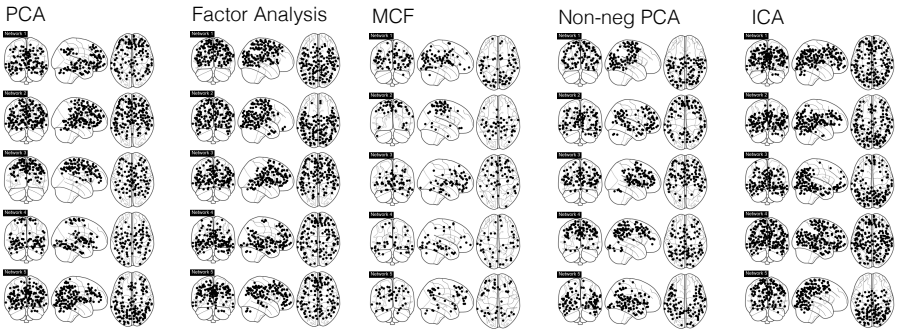

Fig 13. Inferred networks using alternative linear latent variable models. In the case of models such as PCA and factor analysis, networks were obtained by thresholding entries of W so only non-negative entries considered.

S5 Fig. Generalization performance of brain age prediction on HCP and ATR Wide-Age-Range datasets

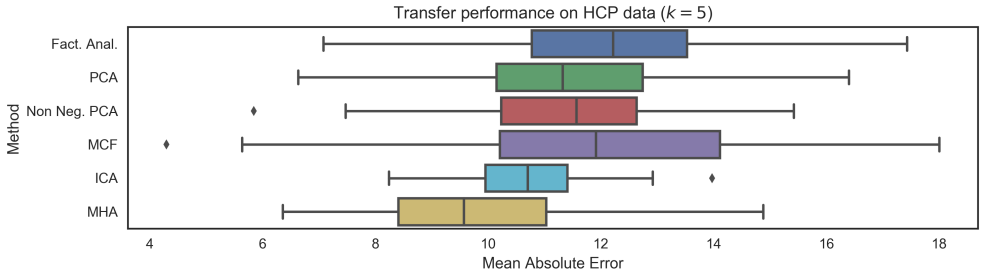

Fig 14. Mean absolute error (MAE) performance on unseen data from HCP repository. Results are broadly consistent with performance on the CamCAN data, indicating good generalization. We note that the introduction of non-negativity or orthogonality constraints leads to improved generalization. The number of functional networks was $k = 5$.

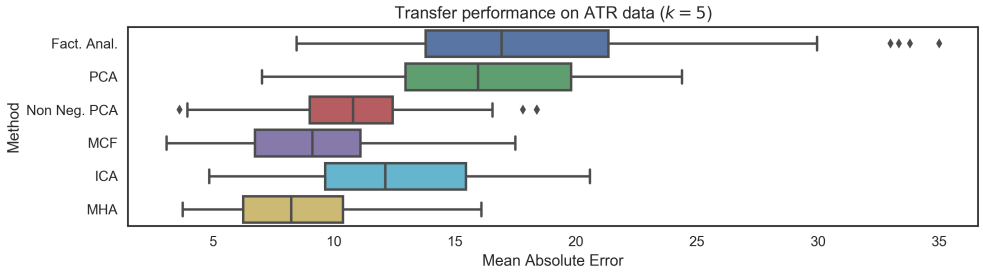

Fig 15. Mean absolute error (MAE) performance on unseen data from ATR Wide-Age-Range repository. Results are broadly consistent with performance on the CamCAN data, indicating good generalization. Further, as with the HCP data, we note that the introduction of non-negativity or orthogonality constraints leads to improved generalization. The number of functional networks considered was $k = 5$.

Acknowledgments

The authors with to thank Steve Smith for valuable feedback and discussions.

References

A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

J. Ashburner. Computational anatomy with the SPM software. *Magnetic Resonance Imaging*, 27(8):1163–1174, 2009.

A. Athreya, D. E. Fishkind, M. Tang, C. E. Priebe, Y. Park, J. T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

C. M. Bennett and M. B. Miller. How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, 1191(1):133–155, 2010.

D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic Press, 2014.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

V. D. Calhoun, R. Miller, G. Pearlson, and T. Adalı. The chronnectome: time-varying connectivity networks as the next frontier in fMRI data discovery. *Neuron*, 84(2): 262–274, 2014.

V. L. Cherkassky, R. K. Kana, T. A. Keller, and M. A. Just. Functional connectivity in a baseline resting-state network in Autism. *Neuroreport*, 17(16):1687–1690, 2006. doi: 10.1097/01.wnr.0000239956.45448.4c.

A. W. Chung, E. Pesce, R. P. Monti, and G. Montana. Classifying hcp task-fMRI networks using heat kernels. In *2016 International Workshop on Pattern Recognition in NeuroImaging (PRNI)*, pages 1–4. IEEE, 2016.

J. H. Cole and K. Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in Neurosciences*, 40(12):681–690, 2017.

J. H. Cole, R. Leech, D. J. Sharp, and A. D. N. Initiative. Prediction of brain age suggests accelerated atrophy after traumatic brain injury. *Annals of Neurology*, 77(4): 571–581, 2015.

J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

J. H. Cole, S. J. Ritchie, M. E. Bastin, M. V. Hernández, S. M. Maniega, N. Royle, J. Corley, A. Pattie, S. E. Harris, Q. Zhang, et al. Brain age predicts mortality. *Molecular Psychiatry*, 23(5):1385, 2018.

H. R. Cremers, T. D. Wager, and T. Yarkoni. The relation between statistical power and inference in fMRI. *PloS one*, 12(11):e0184923, 2017.

J. S. Damoiseaux, K. E. Prater, B. L. Miller, and M. D. Greicius. Functional connectivity tracks clinical deterioration in Alzheimer's disease. *Neurobiology of Aging*, 33(4), 2012. doi: 10.1016/j.neurobiolaging.2011.06.024.

N. U. Dosenbach, B. Nardos, A. L. Cohen, D. A. Fair, J. D. Power, J. A. Church, S. M. Nelson, G. S. Wig, A. C. Vogel, C. N. Lessov-Schlaggar, et al. Prediction of individual brain maturity using fMRI. *Science*, 329(5997):1358–1361, 2010.

F. Esposito, T. Scarabino, A. Hyvärinen, J. Himberg, E. Formisano, S. Comani, G. Tedeschi, R. Goebel, E. Seifritz, and F. Di Salle. Independent component analysis of fMRI group studies by self-organizing clustering. *Neuroimage*, 25(1):193–205, 2005.

K. Franke, G. Ziegler, S. Klöppel, C. Gaser, A. D. N. Initiative, et al. Estimating the age of healthy subjects from t1-weighted MRI scans using kernel methods: exploring the influence of various parameters. *Neuroimage*, 50(3):883–892, 2010.

K. Franke, C. Gaser, B. Manor, and V. Novak. Advanced brainage in older adults with type 2 diabetes mellitus. *Frontiers in Aging Neuroscience*, 5:90, 2013.

L. Friedman, G. H. Glover, F. Consortium, et al. Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. *Neuroimage*, 33(2):471–481, 2006.

L. Geerligs, R. J. Renken, E. Saliasi, N. M. Maurits, and M. M. Lorist. A brain-wide study of age-related changes in functional connectivity. *Cerebral Cortex*, 25(7): 1987–1999, 2014.

L. Geerligs, M. Rubinov, R. N. Henson, et al. State and trait components of functional connectivity: individual differences vary with mental state. *Journal of Neuroscience*, 35(41):13949–13961, 2015.

L. Geerligs, K. A. Tsvetanov, and R. N. Henson. Challenges in measuring individual differences in functional connectivity using fMRI: the case of healthy aging. *Human Brain Mapping*, 38(8):4125–4156, 2017.

L. Geerligs et al. Reduced specificity of functional connectivity in the aging brain during task performance. *Human Brain Mapping*, 35:319–330, 2012.

C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage*, 14(1):21–36, 2001.

C. Grady, S. Sarraf, C. Saverino, and K. Campbell. Age differences in the functional interactions among the default, frontoparietal control, and dorsal attention networks. *Neurobiology of Aging*, 41:159–172, 2016.

H. H. Harman. *Modern Factor Analysis*. Univ. of Chicago Press, 1960.

J. Himberg, A. Hyvärinen, and F. Esposito. Validating the independent components of neuroimaging time series via clustering and visualization. *Neuroimage*, 22(3): 1214–1222, 2004.

J. Hirayama, A. Hyvärinen, V. Kiviniemi, M. Kawanabe, and O. Yamashita. Characterizing variability of modular brain connectivity with constrained principal component analysis. *PloS One*, 11(12):e0168180, 2016.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933.

A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on Neural Networks*, 10(3):626–634, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

A. Hyvärinen, J. I. Hirayama, V. Kiviniemi, and M. Kawanabe. Orthogonal Connectivity Factorization: Interpretable Decomposition of Variability in Correlation Matrices. *Neural Computation*, 28(3):445–484, 2016.

I. Jolliffe. *Principal component analysis*. Springer, 2011.

C. Kelly, B. B. Biswal, R. C. Craddock, F. X. Castellanos, and M. P. Milham. Characterizing variation in the functional connectome: promise and pitfalls. *Trends in Cognitive Sciences*, 16(3):181–188, 2012.

I. Khemakhem, D. P. Kingma, R. P. Monti, and A. Hyvärinen. Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*, 2019.

N. Koutsouleris, C. Davatzikos, S. Borgwardt, C. Gaser, R. Bottlender, T. Frodl, P. Falkai, A. Riecher-Rössler, H.-J. Möller, M. Reiser, et al. Accelerated brain aging in Schizophrenia and beyond: a neuroanatomical marker of psychiatric disorders. *Schizophrenia bulletin*, 40(5):1140–1153, 2013.

J. Lancaster, R. Lorenz, R. Leech, and J. H. Cole. Bayesian optimization for neuroimaging pre-processing in brain age classification and prediction. *Frontiers in Aging Neuroscience*, 10:28, 2018.

N. Leonardi, J. Richiardi, M. Gschwind, S. Simioni, J. M. Annoni, M. Schluep, P. Vuilleumier, and D. Van De Ville. Principal components of functional connectivity: A new approach to study dynamic brain connectivity during rest. *Neuroimage*, 83: 937–950, 2013.

F. Liem, L. Geerligs, J. S. Damoiseaux, and D. S. Margulies. Functional Connectivity in Aging, 2019.

S. Lim, C. E. Han, P. J. Uhlhaas, and M. Kaiser. Preferential detachment during human brain development: age-and sex-specific structural connectivity in diffusion tensor imaging (dti) data. *Cerebral Cortex*, 25(6):1477–1489, 2013.

R. Lorenz, I. R. Violante, R. P. Monti, G. Montana, A. Hampshire, and R. Leech. Dissociating frontoparietal brain networks with neuroadaptive bayesian optimization. *Nature communications*, 9(1):1–14, 2018.

R. Lorenz, L. E. Simmons, R. P. Monti, J. L. Arthur, S. Limal, I. Laakso, R. Leech, and I. R. Violante. Efficiently searching through large tacs parameter spaces using closed-loop bayesian optimization. *Brain stimulation*, 12(6):1484–1489, 2019.

R. Monti, R. Lorenz, P. Hellyer, R. Leech, C. Anagnostopoulos, and G. Montana. Graph embeddings of dynamic functional connectivity reveal discriminative patterns of task engagement in hcp data. In *2015 International Workshop on Pattern Recognition in NeuroImaging*, pages 1–4. IEEE, 2015.

R. P. Monti and A. Hyvärinen. A Unified Probabilistic Model for Learning Latent Factors and Their Connectivities from High-Dimensional Data. In *34th Conference on Uncertainty in Artificial Intelligence*, 2018.

R. P. Monti, P. Hellyer, D. Sharp, R. Leech, C. Anagnostopoulos, and G. Montana. Estimating time-varying brain connectivity networks from functional MRI time series. *NeuroImage*, 103:427–443, 2014.

R. P. Monti, C. Anagnostopoulos, G. Montana, et al. Learning population and subject-specific brain connectivity networks via mixed neighborhood selection. *The Annals of Applied Statistics*, 11(4):2142–2164, 2017a.

R. P. Monti, R. Lorenz, R. M. Braga, C. Anagnostopoulos, R. Leech, and G. Montana. Real-time estimation of dynamic functional connectivity networks. *Human Brain Mapping*, 38(1):202–220, 2017b.

R. P. Monti, R. Lorenz, P. Hellyer, R. Leech, C. Anagnostopoulos, and G. Montana. Decoding time-varying functional connectivity networks via linear graph embedding methods. *Frontiers in Computational Neuroscience*, 11:14, 2017c.

R. P. Monti, K. Zhang, and A. Hyvarinen. Causal discovery with general non-linear relationships using non-linear ica. *arXiv preprint arXiv:1904.09096*, 2019.

T. Ogawa, T. Aihara, T. Shimokawa, and O. Yamashita. Large-scale brain network associated with creative insight: combined voxel-based morphometry and resting-state functional connectivity analyses. *Scientific reports*, 8(1):6477, 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, 2011.

J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, et al. Functional network organization of the human brain. *Neuron*, 72(4):665–678, 2011.

N. Raz and K. M. Rodrigue. Differential aging of the brain: patterns, cognitive correlates and modifiers. *Neuroscience & Biobehavioral Reviews*, 30(6):730–748, 2006.

G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers. *Neuroimage*, 90:449–468, 2014.

H. Sasaki, T. Takenouchi, R. Monti, and A. Hyvärinen. Robust contrastive learning and nonlinear ica in the presence of outliers. *arXiv preprint arXiv:1911.00265*, 2019.

C. D. Sigg and J. M. Buhmann. Expectation-maximization for sparse and non-negative PCA. In *Proceedings of the 25th international conference on Machine learning*, pages 960–967. ACM, 2008.

S. M. Smith. The future of fMRI connectivity. *Neuroimage*, 62(2):1257–1266, 2012.

S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, et al. Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage*, 23:S208–S219, 2004.

S. M. Smith, D. Vidaurre, F. Alfaro-Almagro, T. E. Nichols, and K. L. Miller. Estimation of brain age delta from brain imaging. *NeuroImage*, 2019.

O. Sporns. *Discovering the Human Connectome*. MIT press, 2012.

J. Taylor, N. Williams, R. Cusack, T. Auer, M. Shafto, M. Dixon, L. Tyler, and Cam-CAN. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *NeuroImage*, 18, 2015.

V. G. van de Ven, E. Formisano, D. Prvulovic, C. H. Roeder, and D. E. Linden. Functional connectivity as revealed by spatial independent component analysis of fMRI measurements during rest. *Human Brain Mapping*, 22(3):165–178, 2004.

D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium, et al. The WU-Minn Human Connectome Project: an overview. *Neuroimage*, 80:62–79, 2013.

T. Wu, L. Wang, Y. Chen, C. Zhao, K. Li, and P. Chan. Changes of functional connectivity of the motor network in the resting state in Parkinson's disease. *Neurosci. Lett.*, 460(1):6–10, 2009. doi: 10.1016/j.neulet.2009.05.046.

R. Zass and A. Shashua. Non-negative sparse PCA. In *Advances in Neural Information Processing Systems*, pages 1561–1568, 2007.

A. G. Zippo, I. Castiglioni, J. Lin, V. M. Borsa, M. Valente, and G. E. Biella. Short-term classification learning promotes rapid global improvements of information processing in human brain functional connectome. *Frontiers in Human Neuroscience*, 13:462, 2019a.

A. G. Zippo, V. Del Grosso, A. Patera, M. P. Riccardi, I. G. Tredici, G. Bertoli, A. Mittone, M. Valente, M. Stampanoni, P. Coan, et al. Chronic pain alters microvascular architectural organization of somatosensory cortex. *bioRxiv*, page 755132, 2019b.